

STATISTICAL METHODOLOGY - SAS

INTRODUCTION

The Bureau of the Census conducts the Service Annual Survey (SAS) to provide national estimates of annual revenues, expenses, and e-commerce revenues of establishments in selected service sectors by kind of business.

We develop the estimates in this report using data from a probability sample and administrative records. The sample is taken from a universe of firms operating in the United States and have paid employees. Firms of all sizes are selected in the sample. Administrative records are used to account for firms without paid employees.

This is the first time that e-commerce revenue data were collected in the SAS. The data was collected by adding questions to all SAS questionnaires and mailed to all firms in the sample. The same methods were used to estimate total revenues and e-commerce revenue.

STATISTICAL METHODOLOGY

This section describes the statistical methodology used in the Service Annual Survey.

Sample Design

The Service Annual Survey (SAS) is a probability sample of employer firms engaged in selected service industries. By firm, we mean a business organization consisting of one or more establishments under common ownership or control. (An establishment is a single physical location where business is conducted or where services are performed.) The sample covers both taxable firms and firms exempt from Federal income taxes. Firms with no employees, or nonemployers, are included in the estimates through administrative records data provided by other Federal agencies in the survey year.

Initial Sampling

The sampling frame for SAS was constructed from the Census Bureau's Business Register as of June 1999. The Business Register is a multi-relational database that contains a record for each known establishment connected with an employer firm. A firm can be classified as either a multiunit or a singleunit firm. A multiunit firm is a firm which owns or operates two or more establishments, whereas a singleunit firm is a firm which owns or operates only one establishment. Establishments that are owned by the same multiunit firm are linked using a unique six-digit identification number, called an alpha number, assigned by the Census Bureau. A link between each establishment and its corresponding Employer Identification Number (EIN) is also maintained. The EIN is a number assigned by the Internal Revenue Service (IRS) to any legal entity that intends to hire employees. Under the Federal Insurance Contributions Act

(FICA), each firm with paid employees must have an EIN. The EIN is used by the firm as an identifier to report social security payments for its employees to the IRS.

There is a simple structure that connects an employer firm with its establishments via the EIN. Essentially an employer firm is a cluster of one or more EINs and EINs are clusters of one or more establishments. Each employer firm is associated with at least one EIN and only one firm can use a given EIN. However, an employer firm may use several different EINs for reporting to the IRS. Similarly there is a one-to-many relationship between EINs and establishments. Each EIN is associated with one or more establishments, but each establishment is associated with only one EIN.

The sampling frame for SAS contains two types of sampling units -- alpha numbers and EINs. Both sampling units represent clusters of one or more establishments. The primary stratification of the frame is by kind-of-business group. We further stratify (substratify) the sampling units within kind-of-business groups by a measure of size related to their annual revenue as reported in the 1997 Census of Service Industries. To reduce the variance of the estimates, the sampling units with the largest measures of size are selected "with certainty." This means they are sure to be selected and will represent only themselves (i.e., have a selection probability of one and a sampling weight of one). Within each kind of business a substratum boundary (or cutoff) that divides the certainty units from the noncertainty units is determined. These cutoffs are based on a statistical analysis of data from the 1997 Census of Service Industries. Accordingly, the cutoffs are on a 1997 revenue basis. This analysis is also used to allocate the sample among the kind-of-business groups. The allocation results in an approximate minimum sample size required to achieve desired sampling variability constraints of revenue estimates for specified kind-of-business groups.

The first step in the sample selection identified certainty firms. If a firm had revenue (for 1998 adjusted to a 1997 basis) greater than the certainty cutoff for its major kind of business, the firm was selected into the sample with certainty. For multiunit firms selected with certainty, the sampling unit is the alpha number. For singleunit firms selected with certainty, the sampling unit is the EIN. If a firm was selected with certainty and had more than one establishment at the time of sampling, any new establishments that the firm acquires, even if under new or different EINs, are included in the sample with certainty. This is because the firm was selected using its unique six-digit alpha number. However, if a singleunit firm was selected with certainty, only future establishments associated with that firm's EIN are included with certainty; any new EINs that might later be associated with that firm are subjected to sampling through the quarterly birth-selection procedure (see **Sampling New Employer Firms**).

All firms not selected with certainty were subjected to sampling on an EIN basis. If a firm had more than one EIN, each of its EINs was treated as a separate sampling unit. To be eligible for the initial sampling, a singleunit EIN had to have nonzero payroll in 1998. Multiunit EINs had to have nonzero payroll in 1997 to be eligible for the initial sampling. The EINs were then stratified according to their major kind of business and their estimated 1998 revenue (on a 1997 basis). Within each noncertainty stratum, a simple random sample of EINs was selected. The

sampling rates for the EINs selected from the noncertainty strata varied between 1 in 1.5 and 1 in 1000.

Sampling New Employer Firms (Births)

Periodically, we update the sample to represent new EINs issued by the IRS since the initial sample selection. These EINs, called births, are on the latest available IRS mailing list for FICA taxpayers and may have a kind-of-business classification assigned by the Social Security Administration (SSA).

EIN births are sampled on a quarterly basis (in November of the survey year and in February, May, and August of the year following the survey year) using a two-phase selection procedure. In the first phase, births are stratified by kind of business and a measure of size based on expected employment or quarterly payroll. A relatively large sample is drawn and canvassed to obtain a more reliable measure of size, consisting of revenue in two recent months, and a new or more detailed kind-of-business classification.

Using this more reliable information, the selected births from the first phase are subjected to probability proportional-to-size sampling with overall probabilities equivalent to those used in drawing the initial sample from the Business Register. Because of the time it takes for a new employer firm to acquire an EIN from the IRS, and because of the time needed to accomplish the two-phase birth-selection procedure, EIN births are added to the sample approximately six to nine months after they begin operation.

The EIN births that are selected in the quarterly birth-selection procedure in November of the survey year are included in the initial mailing of the SAS questionnaires in January of the following year.

To better represent all EIN births in the survey year, and specifically to account for the coverage lag in the birth-selection procedure, we update the sample with EIN births that are selected in the year following the survey year. We mail survey forms to these births in June and August to supplement the initial survey mailing.

If a selected EIN ceases to be an employer, it becomes inactive. An inactive EIN is not mailed if it becomes inactive prior to the initial mailout of the survey year. An inactive EIN that resumes its employer status during the survey year is reactivated and mailed during the initial mailing (if active at the time) or as part of one of the two supplemental mailings.

To be eligible for the sample canvass and tabulation in a given year, a noncertainty EIN must meet both of the following requirements:

- ! It must be on the latest available IRS mailing list for FICA taxpayers from the previous quarter.

- ! It must have been selected from the Business Register in either the initial sampling or during the quarterly birth-selection procedure.

EINs selected into the sample with certainty are not dropped from canvass and tabulation if they are no longer on the IRS mailing list. Rather, the firm that used the EIN is contacted, and if a successor EIN is found, it is added to the survey. For both inactive and reactivated EINs, data are tabulated for only the portion of the survey period that these EINs reported payroll to the IRS.

Estimation Procedures

Estimates from this survey are based on the summation of weighted data (reported and imputed), where the weight for a given sampling unit is the inverse of its probability of selection into the sample.

Census Adjustment Procedures

The estimates in this report have been linked to the 1997 Economic Census to reduce sampling error and to allow comparability with the census using the following procedure. Unadjusted estimates of total sales and e-commerce sales were formed by summing weighted sampling unit data (both reported and imputed). Estimated total sales for 1998 for detailed NAICS kinds of businesses were then adjusted to the 1997 Economic Census sales totals. For each detailed kind of business, the ratio of the adjusted 1998 sales to the unadjusted 1998 sales was multiplied by the unadjusted e-commerce sales to derive an adjusted e-commerce estimate. The detailed e-commerce estimates were summed to produce the estimates in this report.

Dollar Values

All dollar values presented are expressed in current dollars; that is, the estimates are not adjusted to a constant dollar series. Consequently, when comparing estimates to prior years, users also should consider price level changes.

Census Disclosure Rules

In accordance with Federal law governing Census Bureau reports, no estimates are published that would disclose the operations of an individual firm.

RELIABILITY OF THE ESTIMATES

An estimate based on a sample survey potentially contains two types of errors - sampling and nonsampling. Sampling error occurs because characteristics differ among sampling units and because only a subset of the entire population is measured in a sample survey. Nonsampling error encompasses all other factors that contribute to the total error of a sample survey estimate. The accuracy of a survey result may be affected by these two types of errors.

Sampling and nonsampling errors are often measured by the quantities, bias and variance. The *bias* of an estimator of an unknown population value is the difference, averaged over all possible samples of the same size and design, between the estimator and the unknown population value. Any systematic error, or inaccuracy that affects all samples of a specified design in a similar way, may bias the resulting estimates. The *variance* of an estimator is the squared difference, averaged over all possible samples of the same size and design, between the estimator and its average value.

Descriptions of sampling and nonsampling errors for the SAS are provided in the following sections.

Sampling Error

Because the estimates are based on a sample, exact agreement with results that would be obtained from a complete enumeration of firms represented on the sampling frame using the same enumeration procedures is not expected. However, because each firm on the sampling frame has a known probability of being selected into the sample, it is possible to estimate the sampling variability of the survey estimates.

The particular sample used in this survey is one of a large number of samples of the same size that could have been selected using the same design. If all possible samples had been surveyed, under the same conditions, an estimate of an unknown population value could have been obtained from each sample. These samples give rise to a distribution of estimates for the unknown population value. A statistical measure of the variability among these estimates is the standard error, which can be approximated from any one sample. The *standard error* is defined as the square root of the variance. The *coefficient of variation* (or relative standard error) of an estimator is the standard error of the estimator divided by the estimator. Note that measures of sampling variability, such as the standard error or coefficient of variation, are estimated from the sample and are also subject to sampling variability. (Technically, we should refer to the *estimated* standard error or the *estimated* coefficient of variation of an *estimator*. However, for the sake of brevity we have omitted this detail.) It is important to note that the standard error and coefficient of variation only measure sampling variability. They do not measure any systematic biases in the estimates. Estimated coefficients of variation for dollar volume estimates and estimated ratios are shown in Table 3A. (All coefficients of variation are expressed as percents.)

The estimate from a particular sample and the approximate standard error associated with the estimate can be used to construct a confidence interval. A *confidence interval* is a range about a given estimator that has a specified probability of containing the estimator's corresponding, unknown population value. Associated with each interval is a percentage of confidence, which is interpreted as follows. If, for each possible sample, an estimate of an unknown population value and its approximate standard error were obtained, then:

1. For approximately 90 percent of the possible samples, the interval from 1.645 standard errors below to 1.645 standard errors above the estimate would include the unknown population value.
2. For approximately 95 percent of the possible samples, the interval from two standard errors below to two standard errors above the estimate would include the unknown population value.

To illustrate the computation of a confidence interval for an estimate of total revenue, assume that an estimate of total revenue is \$10,750 million and the coefficient of variation for this estimate is 1.8 percent, or 0.018. First obtain the standard error of the estimate by multiplying the total revenue estimate by the coefficient of variation. For this example, multiply \$10,750 million by 0.018. This yields a standard error of \$193.5 million. The upper and lower bounds of the 90-percent confidence interval are computed as \$10,750 million plus or minus 1.645 times \$193.5 million. Consequently, the 90-percent confidence interval is \$10,432 million to \$11,068 million. If corresponding confidence intervals were constructed for all possible samples of the same size and design, approximately 9 out of 10 (90 percent) of these intervals would contain the unknown population value.

Nonsampling Error

Nonsampling error encompasses all other factors that contribute to the total error of a sample survey estimate and may also occur in censuses. It is often helpful to think of nonsampling error as arising from deficiencies or mistakes at some point in the survey process. In the SAS, nonsampling error can be attributed to many sources: inability to obtain information about all units in the sample; response errors; differences in the interpretation of the questions; mistakes in coding or keying the data obtained; and other errors of collection, response, coverage, and processing. Although no direct measurement of the potential biases due to nonsampling error has been obtained, precautionary steps were taken in all phases of the collection, processing, and tabulation of the data in an effort to minimize their influence.

A potential source of bias in the estimates is nonresponse. Nonresponse is defined as the inability to obtain all the intended measurements or responses about all selected firms. Two types of nonresponse are often distinguished. *Unit nonresponse* is used to describe the inability to obtain any of the substantive measurements about a sampled firm. In most cases of unit nonresponse, the questionnaire was never returned to the Census Bureau, after several attempts to elicit a response. *Item nonresponse* occurs either when a question is unanswered or the response to the question fails computer or analyst edits.

For both unit and item nonresponse, a missing value is replaced by a predicted value obtained from an appropriate model for nonresponse. This procedure is called *imputation*. Imputed revenue amounts to about 10.9 percent of the total services revenue estimate. Imputed e-commerce revenue accounts for about 22.0 percent of the total e-commerce revenue estimate.